Shortest Repetition-Free Words Accepted by Automata

Hamoon Mousavi and Jeffrey Shallit

School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1 Canada
{sh2mousa,shallit}@uwaterloo.ca

Abstract. We consider the following problem: given that a finite automaton M of N states accepts at least one k-power-free (resp., overlap-free) word, what is the length of the shortest such word accepted? We give upper and lower bounds which, unfortunately, are widely separated.

1 Introduction

Let L be an interesting language, such as the language of primitive words, or the language of non-palindromes. We are interested in the following kind of question: *given that an automaton M of N states accepts a member of L, what is a good bound on the length $\ell(N)$ of the shortest word accepted?*

For example, Ito et al. [7] proved that if L is the language of primitive words, then $\ell(N) \leq 3N - 3$. Horváth et al. [6] proved that if L is the language of non-palindromes, then $\ell(N) \leq 3N$. For additional results along these lines, see [1].

For an integer $k \geq 2$, a *k-power* is a nonempty word of the form x^k. A word is *k-power-free* if it has no k-powers as factors. A word of the form $axaxa$, where a is a single letter, and x is a (possibly empty) word, is called an *overlap*. A word is *overlap-free* if it has no factor that is an overlap.

In this paper we address two open questions left unanswered in [1], corresponding to the case where L is the language of k-power-free (resp., overlap-free) words. For these words and a large enough alphabet we give a class of DFAs of N states for which the shortest k-power (resp., overlap) is of length $N^{\frac{1}{4}(\log N)+O(1)}$. For overlaps over a binary alphabet we give an upper bound of $2^{O(N^{4N})}$.

2 Notation

For a finite alphabet Σ, let Σ^* denote the set of finite words over Σ. Let $w = a_0a_1 \cdots a_{n-1} \in \Sigma^*$ be a word. Let $w[i] = a_i$, and let $w[i..j] = a_i \cdots a_j$. By convention we have $w[i] = \epsilon$ for $i < 0$ or $i \geq n$, and $w[i..j] = \epsilon$ for $i > j$. A prefix p of w is a *period* of w if $w[i + r] = w[i]$ for $0 \leq i < |w| - r$, where $r = |p|$.

For words x, y, let $x \preceq y$ denote that x is a factor of y. A factor x of y is *proper* if $x \neq y$. Let $x \preceq_p y$ (resp., $x \preceq_s y$) denote that x is a prefix (resp., suffix) of y. Let $x \prec_p y$ (resp., $x \prec_s y$) denote that x is a prefix (resp., suffix) of y and $x \neq y$.

A word is *primitive* if it is not a k-power for any $k \geq 2$. Two words x, y are *conjugate* if one is a cyclic shift of the other; that is, if there exist words u, v such that $x = uv$ and $y = vu$. One simple observation is that all conjugates of a k-power are k-powers.

Let $h : \Sigma^* \to \Sigma^*$ be a morphism, and suppose $h(a) = ax$ for some letter a. The *fixed point* of h, starting with $a \in \Sigma$, is denoted by $h^\omega(a) = a\, x\, h(x)\, h^2(x) \cdots$. We say that a morphism h is k-power-free (resp., overlap-free) if $h(w)$ is k-power-free (resp., overlap-free) if w is.

Let $\Sigma_m = \{0, 1, \ldots, m-1\}$. Define the morphism $\mu : \Sigma_2^* \to \Sigma_2^*$ as follows

$$\mu(0) = 01$$
$$\mu(1) = 10.$$

We call $\mathbf{t} = \mu^\omega(0)$ the *Thue-Morse word.* It is easy to see that

$$\mu(\mathbf{t}[0..n-1]) = \mathbf{t}[0..2n-1] \text{ for } n \geq 0.$$

From classical results of Thue [10,11], we know that the morphism μ is overlap-free. From [2], we know that that $\mu(x)$ is k-power free for each $k > 2$.

For a DFA $D = (Q, \Sigma, \delta, q_0, F)$ the set of states, input alphabet, transition function, set of final states, and initial state are denoted by Q, Σ, δ, F, and q_0, respectively. Let $L(D)$ denote the language accepted by D. As usual, we have $\delta(q, wa) = \delta(\delta(q, w), a)$ for a word w.

We state the following basic result without proof.

Proposition 1. *Let $D = (Q, \Sigma, \delta, q_0, F)$ be a (deterministic or nondeterministic) finite automaton. If $L(D) \neq \emptyset$, then D accepts at least one word of length smaller than $|Q|$.*

3 Lower bound

In this section, we construct an infinite family of DFAs such that the shortest k-power-free word accepted is rather long, as a function of the number of states. Up to now only a linear bound was known.

For a word w of length n and $i \geq 1$, let

$$\mathrm{cyc}_i(w) = w[i..n-1]\, w[0..i-2]$$

denote w's ith cyclic shift to the left, followed by removing the last symbol. Also define

$$\mathrm{cyc}_0(w) = w[0..n-2].$$

For example, we have

$$\mathrm{cyc}_2(\texttt{recompute}) = \texttt{computer},$$
$$\mathrm{cyc}_4(\texttt{richly}) = \texttt{lyric}.$$

We call each $\mathrm{cyc}_i(w)$ a *partial conjugate* of w, which is not a reflexive, symmetric, or transitive relation.

A word w is a *simple k-power* if it is a k-power and it contains no k-power as a proper factor.

We start with a few lemmas.

Lemma 2. *Let $w = p^k$ be a simple k-power. Then the word p has $|p|$ distinct conjugates.*

Proof. By contradiction. If p^k is a simple k-power, then p is a primitive word. Suppose that $p = uv = xy$ such that $x \prec_p u$ and $yx = vu$. Without loss of generality, we can assume that $xv \neq \epsilon$. Then there exists a word $t \neq \epsilon$ such that $u = xt$ and $y = tv$. From $vu = yx$ we get

$$vxt = tvx.$$

Using the second theorem of Lyndon and Schützenberger [8], we get that there exists $z \neq \epsilon$ such that

$$vx = z^i$$

$$t = z^j$$

for some positive integers i, j. So $yx = z^{i+j}$, and hence $p = xy$ is not primitive, a contradiction. □

Lemma 3. *Let w be a simple k-power of length n. Then we have*

$$\mathrm{cyc}_i(w) = \mathrm{cyc}_j(w) \text{ iff } i \equiv j \pmod{\frac{n}{k}}. \tag{1}$$

Proof. Let $w = p^k$. If $i \equiv i' \pmod{\frac{n}{k}}$ and $i' < \frac{n}{k}$, then

$$\mathrm{cyc}_i(w) = (p[i'..\frac{n}{k} - 1]\, p[0..i' - 1])^{k-1}\, \mathrm{cyc}_{i'}(p).$$

Similarly, if $j \equiv j' \pmod{\frac{n}{k}}$ and $j' < \frac{n}{k}$, then

$$\mathrm{cyc}_j(w) = (p[j'..\frac{n}{k} - 1]\, p[0..j' - 1])^{k-1}\, \mathrm{cyc}_{j'}(p).$$

If $i' = j'$, then clearly $\mathrm{cyc}_i(w) = \mathrm{cyc}_j(w)$. If $i' \neq j'$, we get that

$$p[i'..\frac{n}{k} - 1]\, p[0..i' - 1] \neq p[j'..\frac{n}{k} - 1]\, p[0..j' - 1]$$

using Lemma 2, and hence $\mathrm{cyc}_i(w) \neq \mathrm{cyc}_j(w)$.

□

Lemma 4. *All conjugates of a simple k-power are simple k-powers.*

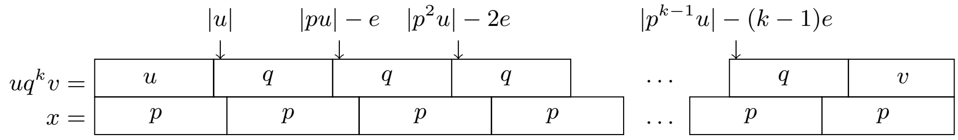


Fig. 1: starting positions of the occurrences of q inside x

Proof. By contradiction. Let $w = p^k$ be a simple k-power, and let $z \neq w$ be a conjugate of w. Clearly z is a k-power. Suppose z contains q^k and $z \neq q^k$. Thus $|q| < |p|$. Since w is simple $q^k \not\preceq w = p^k$. The word $x = p^{k+1}$ contains z as a factor. So $x = uq^kv$, for some words $u, v \preceq p$.
Note that u and v are nonempty and not equal to p since $q^k \not\preceq p^k$. Letting $e := |p| - |q|$, and considering the starting positions of the occurrences of q in x (see Fig. 1), we can write

$$x[|p^iu| - ie..|p^iu| - (i-1)e - 1] = x[|p^ju| - je..|p^ju| - (j-1)e - 1]$$

for every $0 \leq i, j < k$. Since p is a period of x, we can write

$$x[|u| - ie..|u| - (i-1)e - 1] = x[|u| - je..|u| - (j-1)e - 1]$$

which means $x[u - (k-1)e..u + e - 1] \preceq w$ is a k-power. Therefore w contains a k-power other than itself, a contradiction. □

Corollary 5. *Partial conjugates of simple k-powers are k-power-free.*

The next lemma shows that there are infinitely many simple k-powers over a binary alphabet for $k > 2$. We also show that there are infinitely many simple squares over a ternary alphabet, using a result of Currie [4].

Lemma 6.

(i) Let $p = \mathbf{t}[0..2^n - 1]$ where $n \geq 0$. For every $k > 2$, the word p^k is a simple k-power.
(ii) There are infinitely many simple squares over a ternary alphabet.

Proof.

(i) By induction on n. For $n = 0$ we have $p^k = 0^k$ which is a simple k-power. Suppose $n > 0$. To get a contradiction, suppose that there exist words u, v, x with $uv \neq \epsilon$ and $x \neq \epsilon$ such that $p^k = ux^kv$. Note that $|x| < |p|$, so $|uv| \geq k$. Without loss of generality, we can assume that $|v| \geq \lceil \frac{k}{2} \rceil \geq 2$. Let $q = \mathbf{t}[0..2^{n-1} - 1]$. We know that

$$p^k = \mu(q^k).$$

We can write

$$w = ux^k \preceq_p \mu(q^{k-1}q[0..|q| - 2]).$$

Since μ is k-power-free, the word $q^{k-1}q[0..|q| - 2]$ contains a k-power. Hence q^k contains at least two k-powers, a contradiction.

(ii) Currie [4] proved that over a ternary alphabet, for every $n \geq 18$, there is a word p of length n such that all its conjugates are squarefree. Such squarefree words are called *circularly squarefree words*.

We claim that for every circularly squarefree word p, the word p^2 is a simple square. To get a contradiction, let q^2 be the smallest square in p^2. So there exist words u, y with $uy \neq \epsilon$ such that $p^2 = uq^2y$. We have $|q^2| > |p|$ since p is circularly squarefree. Therefore, if we let $p = uv = xy$, then $|x| > |u|$ and $|v| > |y|$. So there exists t such that $x = ut$ and $v = ty$. We can assume $|t| < |q|$, since otherwise $|t| = |q|$ and $|uy| = 0$, a contradiction. Now since $q^2 = vx = tyut$, we get that q begins and ends with t, which means $t^2 \prec q^2$. Therefore p^2 has a smaller square than q^2, a contradiction.

□

Next we show how to construct arbitrarily long simple k-powers from smaller ones. Fix $k = 2$ (resp., $k \geq 3$) and $m = 3$ (resp., $m = 2$). Let $w_1 \in \Sigma_m^*$ be a simple k-power. Using the previous lemma, there are infinitely many choices for w_1. Let w_1 be of length n. Define $w_{i+1} \in \Sigma_{m+i}^*$ for $i \geq 1$ recursively by

$$w_{i+1} = \mathrm{cyc}_0(w_i)a_i\,\mathrm{cyc}_{n^{i-1}}(w_i)a_i\,\mathrm{cyc}_{2n^{i-1}}(w_i)a_i \cdots \mathrm{cyc}_{(n-1)n^{i-1}}(w_i)a_i \tag{2}$$

where $a_i = m + i - 1$ and $w_0 = 0$. The next lemma states that w_i, for $i \geq 1$, is a simple k-power. Therefore, using Corollary 5, each word $\mathrm{cyc}_0(w_i)$ is k-power-free. For $i \geq 1$, it is easy to see that

$$|w_i| = n|w_{i-1}| = n^i. \tag{3}$$

Lemma 7. *For every $i \geq 1$, the word w_i is a simple k-power.*

Proof. By induction on i. The word w_1 is a simple k-power. Now suppose that w_i is a simple k-power for some $i \geq 1$. Using Lemma 3, we have $\mathrm{cyc}_{jn^{i-1}}(w_i) = \mathrm{cyc}_{(j+\frac{n}{k})n^{i-1}}(w_i)$, since $\frac{|w_i|}{k} = \frac{n^i}{k}$.

We now claim that w_{i+1} is a k-power and

$$w_{i+1} = (\mathrm{cyc}_0(w_i)a_i\,\mathrm{cyc}_{n^{i-1}}(w_i)a_i\,\mathrm{cyc}_{2n^{i-1}}(w_i)a_i \cdots \mathrm{cyc}_{(\frac{n}{k}-1)n^{i-1}}(w_i)a_i)^k.$$

To see this, suppose that w_{i+1} contains a k-power y^k such that $w_{i+1} \neq y^k$.

If y contains more than one occurrence of a_i, then $y = ua_i\,\mathrm{cyc}_j(w_i)a_iv$ for some words u, v and an integer j. Since $y^2 = ua_i\,\mathrm{cyc}_j(w_i)a_ivua_i\,\mathrm{cyc}_j(w_i)a_iv \preceq w_{i+1}$, using (2) and Lemma 3, we get

$$|y| = \left|\mathrm{cyc}_j(w_i)a_ivua_i\right| \geq \frac{n}{k}n^i = \frac{|w_{i+1}|}{k},$$

and hence $y^k = w_{i+1}$, a contradiction.

If y contains just one a_i, then $y = ua_iv$ for some words u, v which contain no a_i. So $y^k = u(avu)^{k-1}av$ for $a = a_i$. Therefore vu is a partial conjugate of w_i. However the distance between two equal partial conjugates of w_i in w_{i+1} is longer than just one letter, using (2) and Lemma 3.

Finally, if y contains no a_i, then a partial conjugate of w_i contains a k-power, which is impossible due to Lemma 4. □

To make our formulas easier to read, we define $a_0 = w_1[n-1]$.

Theorem 8. *For $i \geq 1$, there is a DFA D_i with $2^{i-1}(n-1)+2$ states such that $\mathrm{cyc}_0(w_i)$ is the shortest k-power-free word in $L(D_i)$.*

Proof. Define $D_1 = (Q_1, \Sigma_{a_1}, \delta_1, q_{1,0}, F_1)$ where

$$\begin{aligned}
&Q_1 := \{q_{1,0}, q_{1,1}, q_{1,2}, \ldots, q_{1,n-1}, q_d\},\\
&F_1 := \{q_{1,n-1}\},\\
&\delta_1(q_{1,j}, w[j]) := q_{1,j+1} \text{ for } 0 \leq j < n-1,
\end{aligned}$$

and the rest of the transitions go to the dead state q_d. Clearly we have $|Q_1| = n+1$ and $L(D_1) = \{\mathrm{cyc}_0(w_1)\}$.

We define $D_i = (Q_i, \Sigma_{a_i}, \delta_i, q_{1,0}, F_i)$ for $i \geq 2$ recursively. For the rest of the proof s and t denote (possibly empty) sequences of integers and j denotes a single integer (a sequence of length 1). We use integer sequences as subscripts of states in Q_i. For example, $q_{1,0}$, $q_{s,j}$, and $q_{s,2,t}$ might denote states of D_i. For $i \geq 1$, define

$$\begin{aligned}
&Q_{i+1} := Q_i \cup \{q_{i+1,t} : q_t \in (Q_i - F_i) - \{q_d\}\}, && (4)\\
&F_{i+1} := \{q_{i+1,i,t} : \delta_i(q_{i,t}, c) = q_{1,n-1} \text{ for some } c \in \Sigma_{a_i}\}, && (5)\\
&\text{if } q_t \in Q_i \text{ and } c \in \Sigma_{a_i}, \text{ then } \delta_{i+1}(q_t, c) := \delta_i(q_t, c) && (6)\\
&\text{if } q_t, q_s \in (Q_i - F_i) - \{q_d\}, c \in \Sigma_{a_i}, \text{ and } \delta_i(q_t, c) = q_s,\\
&\qquad\qquad \text{then } \delta_{i+1}(q_{i+1,t}, c) := q_{i+1,s} && (7)\\
&\text{if } q_t \in F_i, \text{ then } \delta_{i+1}(q_t, a_i) := q_{1,1} \text{ and } \delta_{i+1}(q_t, a_{i-1}) := q_{i+1,1,0} && (8)\\
&\text{if } i > 1, q_{i+1,t} \notin F_{i+1}, \text{ and } \delta_i(q_t, a_{i-1}) = q_{1,j},\\
&\qquad\qquad \text{then } \delta_{i+1}(q_{i+1,t}, a_i) := q_{1,j+1} && (9)
\end{aligned}$$

and finally for the special case of $i = 1$,

$$\delta_2(q_{2,1,j}, a_1) := q_{1,j+2} \text{ for } 0 \leq j < n-2. \tag{10}$$

The rest of the transitions, not indicated in (6)–(10), go to the dead state q_d. Fig. 2b depicts D_2 and D_3. Using (4), we have $|Q_{i+1}| = 2|Q_i| - 2 = 2^i(n-1)+2$ by a simple induction.

An easy induction on i proves that $|F_i| = 1$. So let f_i be the appropriate integer sequence for which $F_i = \{q_{f_i}\}$. Using (6)–(10), we get that for every $1 \leq j < n$, there exists exactly one state $q_t \in Q_i$ for which $\delta_i(q_t, a_{i-1}) = q_{1,j}$.

By induction on i, we prove that for $i \geq 2$ if $\delta_i(q_t, a_{i-1}) = q_{1,j}$, then

$$\begin{aligned}
&x_1 = \mathrm{cyc}_{(j-1)n^{i-2}}(w_{i-1}), && (11)\\
&x_2 = w_i[0..jn^{i-1} - 2], && (12)\\
&x_3 = w_i[(j-1)n^{i-1}..n^i - 2]. && (13)
\end{aligned}$$

are the shortest k-power-free words for which

$$\delta_i(q_{1,j-1}, x_1) = q_t, \tag{14}$$
$$\delta_i(q_{1,0}, x_2) = q_t, \tag{15}$$
$$\delta_i(q_{1,j-1}, x_3) = q_{f_i}. \tag{16}$$

In particular, from (13) and (16), for $j = 1$, we get that $\mathrm{cyc}_0(w_i)$ is the shortest k-power-free word in $L(D_i)$.

The fact that our choices of x_1, x_2, and x_3 are k-power-free follows from the fact that proper factors of simple k-powers are k-power-free. For $i = 2$ the proofs of (14)–(16) are easy and left to the readers.

Suppose that (14)–(16) hold for some $i \geq 2$. Let us prove (14)–(16) for $i + 1$. Suppose that

$$\delta_{i+1}(q_t, a_i) = q_{1,j}. \tag{17}$$

First we prove that the shortest k-power-free word x for which

$$\delta_{i+1}(q_{1,j-1}, x) = q_t,$$

is $x = \mathrm{cyc}_{(j-1)n^{i-1}}(w_i)$.

If $q_t \in Q_i$, from (8) and (17), we have

$$q_t = q_{f_i}, \text{ and}$$
$$\delta_{i+1}(q_t, a_i) = q_{1,1}.$$

By induction hypothesis, the $\mathrm{cyc}_0(w_i)$ is the shortest k-power-free word in $L(D_i)$. In other words, we have $\delta_i(q_{1,0}, \mathrm{cyc}_0(w_i)) = q_{f_i} = q_t$, which can be rewritten using (6) as $\delta_{i+1}(q_{1,0}, \mathrm{cyc}_0(w_i)) = q_t$.

Now suppose $q_t \notin Q_i$. Then by (9) and (17), we get that there exists t' such that

$$t = i + 1, t';$$
$$\delta_i(q_{t'}, a_{i-1}) = q_{1,j-1}.$$

From the induction hypothesis, i.e., (15) and (16), we can write

$$\delta_i(q_{1,0}, w_i[0..(j-1)n^{i-1} - 2]) = q_{t'}, \tag{18}$$
$$\delta_i(q_{1,j-1}, w_i[(j-1)n^{i-1}..n^i - 2]) = q_{f_i}. \tag{19}$$

In addition $w_i[0..(j-1)n^{i-1} - 2]$ and $w_i[(j-1)n^{i-1}..n^i - 2]$ are the shortest k-power-free transitions from $q_{1,0}$ to $q_{t'}$ and from $q_{1,j-1}$ to q_{f_i} respectively. Using (6), we can rewrite (18) and (19) for δ_{i+1} as follows:

$$\delta_{i+1}(q_{1,0}, w_i[0..(j-1)n^{i-1} - 2]) = q_{t'}, \tag{20}$$
$$\delta_{i+1}(q_{1,j-1}, w_i[(j-1)n^{i-1}..n^i - 2]) = q_{f_i}. \tag{21}$$

Note that from (7) and (20), we get

$$\delta_{i+1}(q_{i+1,1,0}, w_i[0..(j-1)n^{i-1} - 2]) = q_{i+1,t'} = q_t. \tag{22}$$

We also have $\delta_{i+1}(q_{f_i}, a_i) = q_{i+1,1,0}$, using (8). So together with (21) and (22), we get

$$\delta_{i+1}(q_{1,j-1}, \mathrm{cyc}_{(j-1)n^{i-1}}(w_i)) = q_t$$

and $\mathrm{cyc}_{(j-1)n^{i-1}}(w_i)$ is the shortest k-power-free transition from $q_{1,j-1}$ to q_t.

The proofs of (15) and (16) are similar. □

In what follows, all logarithms are to the base 2.

Corollary 9. *For infinitely many N, there exists a* DFA *with N states such that the shortest k-power-free word accepted is of length $N^{\frac{1}{4}\log N+O(1)}$.*

Proof. Let $i = \lfloor \log n \rfloor$ in Theorem 8. Then $D = D_i$ has

$$N = 2^{\lfloor \log n \rfloor - 1}(n-1) + 2 = \Omega(n^2)$$

states. In addition, the shortest k-power-free word in $L(D)$ is $\mathrm{cyc}_0\left(w_{\lfloor \log n \rfloor}\right)$. Now, using (3) we can write

$$\left|\mathrm{cyc}_0(w_{\lfloor \log n \rfloor})\right| = n^{\lfloor \log n \rfloor} - 1.$$

Suppose $2^t \le n < 2^{t+1} - 1$, so that $t = \lfloor \log n \rfloor$ and Then $\log N = 2t + O(1)$, so $\frac{1}{4}(\log N)^2 = t^2 + O(t)$. On the other hand $\log |w| = \lfloor \log n \rfloor (\log n) = t(t+O(1)) = t^2 + O(t)$. Now $2^{O(t)} = n^{O(1)} = N^{O(1)}$, and the result follows. □

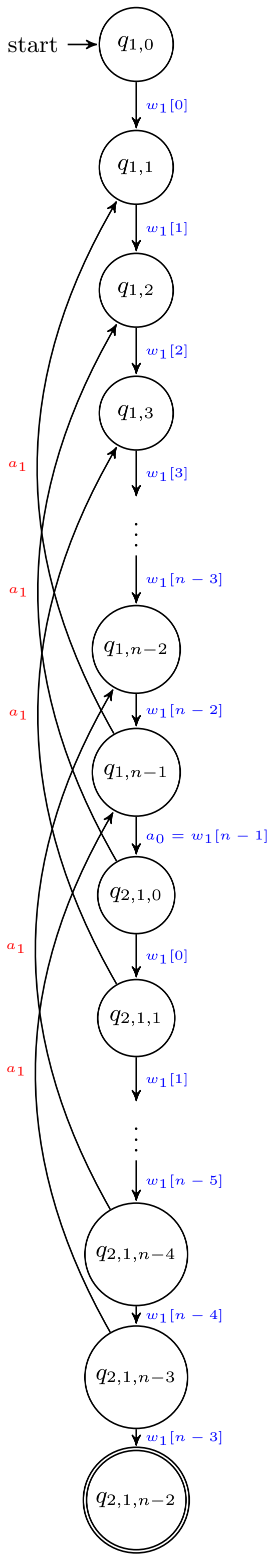
Remark 10. The same bound holds for overlap-free words. To do so, we define a *simple overlap* as a word of the form $axaxa$ where $axax$ is a simple square. In our construction of the DFAs, we use complete conjugates of $(ax)^2$ instead of partial conjugates.

Remark 11. The D_i in Theorem 8 are defined over the growing alphabet Σ_{m+i-1}. However, we can fix the alphabet to be Σ_{m+1}. For this purpose, we introduce w'_i which is quite similar to w_i:

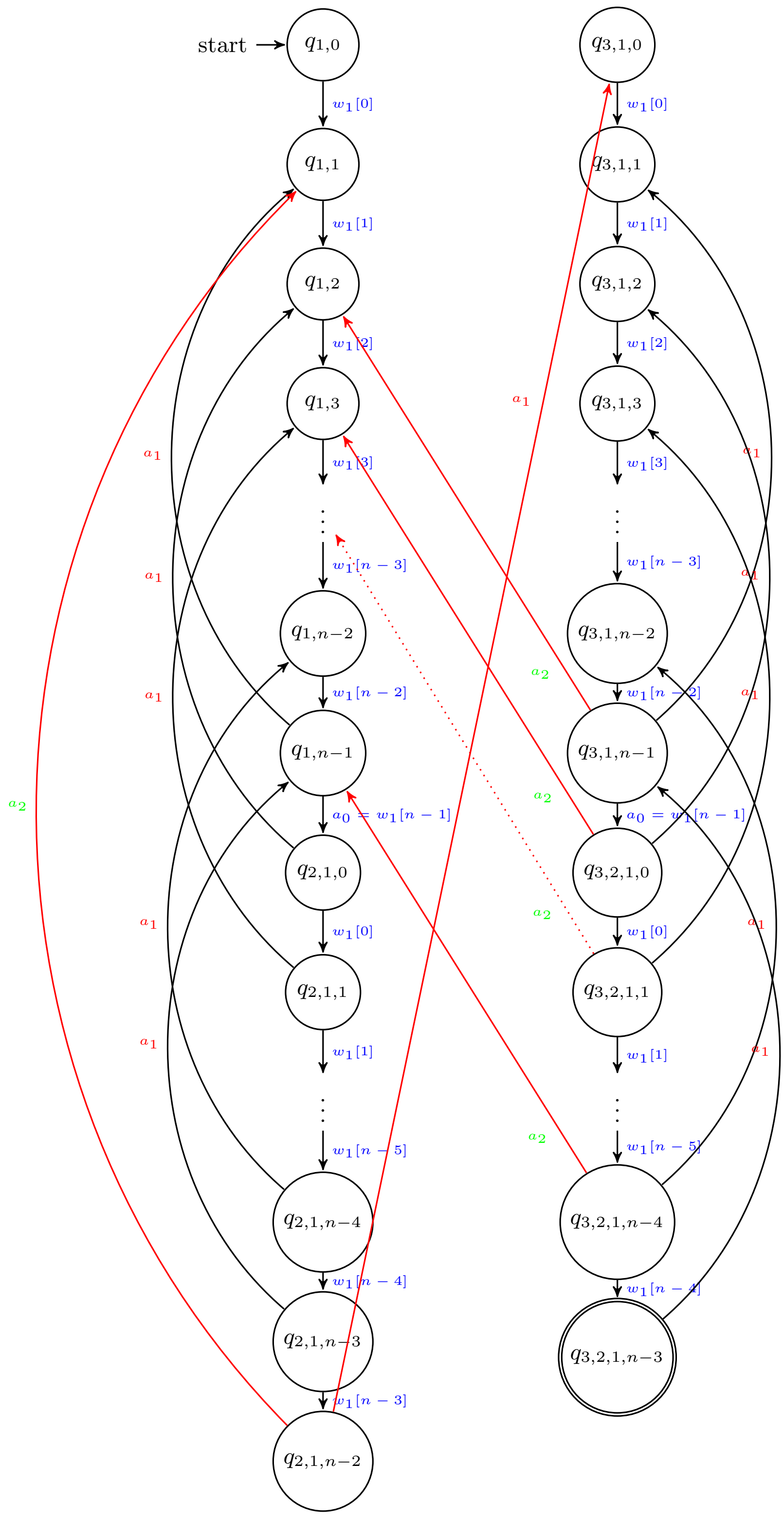
$$\begin{aligned} w'_1 &= w_1, \\ w'_{i+1} &= \mathrm{cyc}_0(w'_i) b_i \, \mathrm{cyc}_{n^{i-1}}(w'_i) b_i \, \mathrm{cyc}_{2n^{i-1}}(w'_i) b_i \cdots \mathrm{cyc}_{(n-1)n^{i-1}}(w'_i) b_i, \end{aligned}$$

where $b_i = mc_im$ such that c_i is (any of) the shortest nonempty k-power-free word over Σ_m not equal to $c_1, \ldots, c_{i-1}$. Clearly we have $|b_i| \le |b_{i-1}| + 1 = O(i)$, and hence $w'_i = \Theta(n^i)$.

One can then prove Lemma 7 and Theorem 8 for w'_i with minor modifications of the argument above. In particular, we construct DFA D'_i that accepts $\mathrm{cyc}_0(w'_i)$ as the shortest k-power-free word accepted, and a D'_i that is quite similar to D_i. In particular, they have asymptotically the same number of states.



(a) transition diagram of D_2



(b) transition diagram of D_3

Fig. 2: transition diagrams

4 Upper bound for overlap-free words

In this section, we prove an upper bound on the length of the shortest overlap-free word accepted by a DFA D over a binary alphabet.

Let $L = L(D)$ and let R be the set of overlap-free words over Σ_2^*. Carpi [3] defined a certain operation Ψ on binary languages, and proved that $\Psi(R)$ is regular. We prove that $\Psi(L)$ is also regular, and hence $\Psi(L) \cap \Psi(R)$ is regular. The next step is to apply Proposition 1 to get an upper bound on the length of the shortest word in $\Psi(L) \cap \Psi(R)$. This bound then gives us an upper bound on the length of the shortest overlap-free word in L.

Let $H = \{\epsilon, 0, 1, 00, 11\}$. Carpi defines maps

$$\Phi_l, \Phi_r : \Sigma_{25} \to H$$

such that for every pair $h, h' \in H$, one has

$$h = \Phi_l(a), h' = \Phi_r(a)$$

for exactly one letter $a \in \Sigma_{25}$.

For every word $w \in \Sigma_{25}^*$, define $\Phi(w) \in \Sigma_2^*$ inductively by

$$\Phi(\epsilon) = \epsilon, \Phi(aw) = \Phi_l(a)\mu(\Phi(w))\Phi_r(a) \qquad (w \in \Sigma_{25}^*, a \in \Sigma_{25}). \tag{23}$$

Expanding (23) for $w = a_0 a_1 \cdots a_{n-1}$, we get

$$\Phi_l(a_0)\mu(\Phi_l(a_1)) \cdots \mu^{n-1}(\Phi_l(a_{n-1}))\mu^{n-1}(\Phi_r(a_{n-1})) \cdots \mu(\Phi_r(a_1))\Phi_r(a_0). \tag{24}$$

For $L \subseteq \Sigma_2^*$ define $\Psi(L) = \bigcup_{x \in L} \Phi^{-1}(x)$. Based on the decomposition of Restivo and Salemi [9] for finite overlap-free words, the language $\Psi(x)$ is always nonempty for an overlap-free word $x \in \Sigma_2^*$. The next theorem is due to Carpi [3].

Theorem 12. *$\Psi(R)$ is regular.*

Carpi constructed a DFA A with less than 400 states that accepts $\Psi(R)$. We prove that Ψ preserves regular languages.

Theorem 13. *Let $D = (Q, \Sigma_2, \delta, q_0, F)$ be a* DFA *with N states, and let $L = L(D)$. Then $\Psi(L)$ is regular and is accepted by a* DFA *with at most N^{4N} states.*

Proof. Let $\iota : Q \to Q$ denote the identity function, and define $\eta_0, \eta_1 : Q \to Q$ as follows

$$\eta_i(q) = \delta(q, i) \text{ for } i = 0, 1. \tag{25}$$

For functions $\zeta_0, \zeta_1 : Q \to Q$, and a word $x = b_0 b_1 \cdots b_{n-1} \in \Sigma_2^*$, define $\zeta_x = \zeta_{b_{n-1}} \circ \cdots \circ \zeta_{b_1} \circ \zeta_{b_0}$. Therefore we have $\zeta_y \circ \zeta_x = \zeta_{xy}$. Also by convention $\zeta_\epsilon = \iota$. So for example $x \in L(D)$ if and only if $\eta_x(q_0) \in F$.

We create DFA $D' = (Q', \Sigma_{25}, \delta', q'_0, F')$ where

$$Q' = \{[\kappa, \lambda, \zeta_0, \zeta_1] : \kappa, \lambda, \zeta_0, \zeta_1 : Q \to Q\},$$
$$\delta'([\kappa, \lambda, \zeta_0, \zeta_1], a) = \left[\zeta_{\Phi_l(a)} \circ \kappa, \lambda \circ \zeta_{\Phi_r(a)}, \zeta_1 \circ \zeta_0, \zeta_0 \circ \zeta_1\right].$$

Also let

$$q'_0 = [\iota, \iota, \eta_0, \eta_1],$$
$$F' = \{[\kappa, \lambda, \zeta_0, \zeta_1] : \lambda \circ \kappa(q_0) \in F\}. \tag{26}$$

We can see that $|Q'| = N^{4N}$. We claim that D' accepts $\Psi(L)$. Indeed, on input w, the DFA D' simulates the behavior of D on $\Phi(w)$.

Let $w = a_0 a_1 \cdots a_{n-1} \in \Sigma_{25}^*$, and define

$$\Phi_1(w) = \Phi_l(a_{a_0})\mu(\Phi_l(a_1)) \cdots \mu^{n-1}(\Phi_l(a_{n-1})),$$
$$\Phi_2(w) = \mu^{n-1}(\Phi_r(a_{n-1})) \cdots \mu(\Phi_r(a_1))\Phi_r(a_0).$$

Using (24), we can write

$$\Phi(w) = \Phi_1(w)\Phi_2(w).$$

We prove by induction on n that

$$\delta'(q'_0, w) = \left[\eta_{\Phi_1(w)}, \eta_{\Phi_2(w)}, \eta_{\mu^n(0)}, \eta_{\mu^n(1)}\right]. \tag{27}$$

For $n = 0$, we have $\Phi(w) = \Phi_1(w) = \Phi_2(w) = \epsilon$. So

$$\delta'(q'_0, \epsilon) = q'_0 = [\iota, \iota, \eta_0, \eta_1] = [\eta_{\Phi_1(w)}, \eta_{\Phi_2(w)}, \eta_{\mu^0(0)}, \eta_{\mu^0(1)}].$$

So we can assume (27) holds for some $n \geq 0$. Now suppose $w = a_0 a_1 \cdots a_n$ and write

$$\begin{aligned}
&\delta'(q'_0, a_0 a_1 \cdots a_n)\\
&= \delta'(\delta'(q'_0, a_0 a_1 \cdots a_{n-1}), a_n)\\
&= \delta'\left(\left[\eta_{\Phi_1(w[0..n-1])}, \eta_{\Phi_2(w[0..n-1])}, \eta_{\mu^n(0)}, \eta_{\mu^n(1)}\right], a_n\right)\\
&= \left[\eta_{\mu^n(\phi_l(a_n))} \circ \eta_{\Phi_1(w[0..n-1])}, \eta_{\Phi_2(w[0..n-1])} \circ \eta_{\mu^n(\phi_r(a_n))}, \eta_{\mu^n(1)} \circ \eta_{\mu^n(0)}, \eta_{\mu^n(0)} \circ \eta_{\mu^n(1)}\right]\\
&= \left[\eta_{\Phi_1(w)}, \eta_{\Phi_2(w)}, \eta_{\mu^{n+1}(0)}, \eta_{\mu^{n+1}(1)}\right],
\end{aligned} \tag{28}$$

and equality (28) holds because

$$\begin{aligned}
&\Phi_1(w[0..n-1])\mu^n(\phi_l(a_n)) = \Phi_1(w),\\
&\mu^n(\phi_r(a_n))\Phi_2(w[0..n-1]) = \Phi_2(w),\\
&\mu^n(0)\mu^n(1) = \mu^n(01) = \mu^n(\mu(0)) = \mu^{n+1}(0), \text{ and similarly}\\
&\mu^n(1)\mu^n(0) = \mu^{n+1}(1).
\end{aligned}$$

Finally, using (26), we have

$$\begin{aligned}
w \in L(D') &\iff \delta'(q'_0, w) = \left[\eta_{\Phi_1(w)}, \eta_{\Phi_2(w)}, \zeta_0, \zeta_1\right] \in F'\\
&\iff \eta_{\Phi_1(w)} \circ \eta_{\Phi_2(w)}(q_0) \in F\\
&\iff \Phi(w) = \Phi_1(w)\Phi_2(w) \in L(D).
\end{aligned}$$

□

Theorem 14. *Let $D = (Q, \Sigma_2, \delta, q_0, F)$ be a* DFA *with N states. If D accepts at least one overlap-free word, then the length of the shortest overlap-free word accepted is $2^{O(N^{4N})}$.*

Proof. Let $L = L(D)$. Using Theorem 13, there exists a DFA D' with N^{4N} states that accepts the language $\Psi(L)$.

Since $\Psi(R)$ is regular and is accepted by a DFA with at most 400 states, we see that

$$K = \Psi(L) \cap \Psi(R)$$

is regular and is accepted by a DFA with $O\left(N^{4N}\right)$ states.

Since L accepts an overlap-free word, the language K is nonempty. Using Proposition 1, we see that K contains a word w of length $O\left(N^{4N}\right)$.

Therefore $\Phi(w)$ is an overlap-free word in L. By induction, one can easily prove that $|\Phi(w)| = O\left(2^{|w|}\right)$. Hence we have $|\Phi(w)| = 2^{O(N^{4N})}$. □

References

1. T. Anderson, J. Loftus, N. Rampersad, N. Santean, and J. Shallit. Detecting palindromes, patterns and borders in regular languages. *Info. Comput.* **207** (2009), 1096-1118.
2. F.-J. Brandenburg. Uniformly growing k-th power-free homomorphisms. *Theoret. Comput. Sci.* **23** (1983), 69–82.
3. A. Carpi. Overlap-free words and finite automata. *Theoret. Comput. Sci.* **115** 1993, 243–260.
4. J. Currie. There are ternary circular square-free words of length n for $n \geq 18$. *Electron. J. Comb.* **9**(1) (2002), Paper #N10. Available at `http://www.combinatorics.org/ojs/index.php/eljc/article/view/v9i1n10`.
5. T. Harju. On cyclically overlap-free words in binary alphabets. In G. Rozenberg and A. Salomaa, eds., *The Book of L*, Springer-Verlag, 1986, pp. 125–130.
6. S. Horváth, J. Karhumäki, and J. Kleijn. Results concerning palindromicity. *J. Info. Process. Cybern. EIK* **23** (1987), 441–451.
7. M. Ito, M. Katsura, H. J. Shyr, and S. S. Yu. Automata accepting primitive words. *Semigroup Forum* **37** (1988), 45–52.
8. R. C. Lyndon and M. P. Schützenberger. The equation $a^M = b^N c^P$ in a free group. *Michigan Math. J.* **9** (1962), 289–298.
9. A. Restivo and S. Salemi. On weakly square-free words. *Bull. EATCS* **21** (1983), 49-56.
10. A. Thue. Über unendliche Zeichenreihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **7** (1906), 1–22. Reprinted in T. Nagell, ed., *Selected Mathematical Papers of Axel Thue*, Universitetsforlaget, Oslo, 1977, pp. 139–158.
11. A. Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichen reihen. *Norske vid. Selsk. Skr. Mat. Nat. Kl.* **1** (1912), 1–67. Reprinted in T. Nagell, ed., *Selected Mathematical Papers of Axel Thue,* Universitetsforlaget, Oslo, 1977, pp. 413–478.